2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Increasing the Convergence Domain of RGB-D Direct Registration Methods for Vision-based Localization in Large Scale Environments

Renato Martins and Patrick Rives

***Abstract*— Developing autonomous vehicles capable of dealing with complex and dynamic unstructured environments over large-scale distances, remains a challenging goal. One of the major difficulties in this objective is the precise localization of the vehicle within its environment so that autonomous navigation techniques can be employed. In this context, this paper presents a methodology to map building and to efficient pose computation which is specially adapted for cases of large displacements. Our method uses hybrid robust RGB-D cost functions that have different convergence properties, whilst exploiting the visibility rotation invariance given by panoramic spherical images. The proposed registration model is composed of a RGB and point-to-plane ICP cost in a multi-resolution framework. We close up the paper presenting mapping and localization results in real outdoor scenes.**

I. INTRODUCTION

Developing autonomous vehicles capable of dealing with complex and dynamic unstructured environments over large-scale distances, remains a challenging goal. One of the major difficulties in this objective is the precise localisation of the vehicle within its environment so that autonomous navigation techniques can be employed. Indeed robust localisation, particularly in heavily occluded areas (such as canyons, tunnels or forest areas), is a non-trivial problem due to GPS masking and poor precision of low cost inertial measurements units (IMU). Classical dead reckoning methods such as odometry, typically performed by inertial sensors and wheels encoders, are prone to drift and therefore are not suited to large distances. Relying on recent advances in sensor technology and on a better understanding of the Intelligent Vehicles requirements, vision based methods sound quite promising to tackle autonomous navigation issues efficiently. Visual odometry (VO), which estimates the full 6 degrees of freedom (DOF) of the vehicle motion from image sequences, produces very precise results and has lower drift than the most expensive IMU's [How08]. Yet, VO methods [NNB04], [CMR10], [KA08] are incremental and then prone to drift when integrated over time.

In [MCR15], we have designed a novel panoramic stereo sensor able to acquire high resolution spherical RGB-D images in real time. Thanks to an offline processing phase, the environment is modeled by a graph of spherical RGB-D keyframes precisely positioned using visual odometry (see fig. 1). This approach, along with the work of [CMR10], allowed direct methods to perform accurate and robust visual odometry.

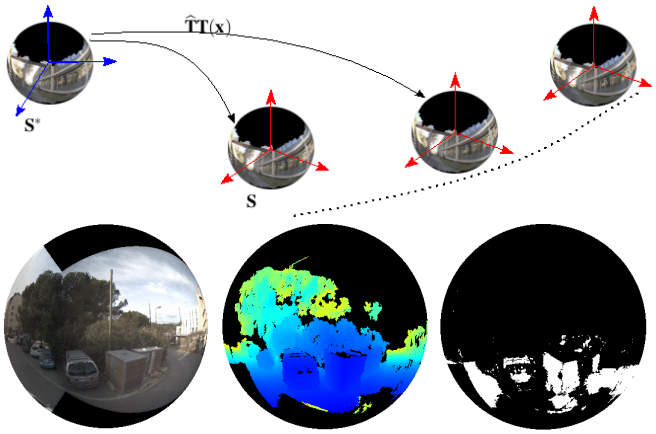

Fig. 1: A keyframe with the RGB spherical image (left), depth (middle) and pixels with higher normal vector confidence in white (right). The confidence index of less stable image regions (e.g. vegetation and no-structured areas) is lower than in structured regions with smooth surfaces.

The following steps were to ensure that the strategy in [MCR15] generates valid and useful long-term models over large-scales. This is important since the validity of the topometric representation (the metric map using a graph topological structure [MC13]) is tightly related to the registration/navigation algorithms that will explore these representations in a posteriori task. Even though a general mathematical condition cannot be established for the convergence, some effort was done in estimating "confidence" navigation envelops by using a teach and repeat paradigm [FB10] [CTG+15]. In this context, we have proposed different strategies to increase the convergence domain as: i) an uncertainty modelling and point stability index based on the visibility of features between successive frames [GMRD15]; ii) the improvement of the information retained in the keyframes by regularising and combining near frames [MFMR15] and; iii) the design of more efficient direct registration methodologies with better convergence properties than classic RGB-D VO methods [MFMR16].

This work addresses particularly the last topic and we show the outcome of a more stable and robust direct registration task in the density/sparsity of the representation (the number of keyframes). Our contribution is on considering, in the registration task, the information gathered from

Inria Sophia Antipolis, France
Email: renato-jose.martins@inria.fr
patrick.rives@inria.fr

This work was supported by CNPq of Brazil under grant number 216026/2013-0.

geometric and photometric error direct cost functions, not only for improving ranking conditioning as in [TAC11] and [KSC13], but for taking into consideration the convexity of both terms to achieve a larger convergence domain and a smaller number of iterations [MFMR16]. This approach is performed within a multi-resolution framework to reduce the computational cost whilst preserving the precision. Most importantly, according to numerical tests, this formulation results in a larger region of attraction and faster convergence than classical RGB, ICP and RGB-D costs as [TAC11]. A direct outcome is the creation of sparser scene models.

The remainder is organized as follows. First, we review the scene model and the basic representation in Section II. Next, we introduce our dense registration methodology in Section III. Lastly, we present experimental mapping results in Section IV for real outdoor contexts, and to conclude the paper in Section V.

II. Preliminaries and Sensor Projection Model

A frame $\mathcal{F} = \{\mathcal{I}, \mathcal{D}\}$ is composed of an image $\mathcal{I} \in [0, 255]^{m \times n}$ as pixel intensities and $\mathcal{D} \in \mathbb{R}_+^{m \times n}$ as the depth. The mapping between the image pixel coordinates $\mathbf{p} \in \mathbb{P}^2$ and depth to 3D Cartesian coordinates is given by the sensor projection $g : \mathbb{P}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}^3$. The sensor projection model of interest here is the spherical (the images are projected in the unit sphere $\mathbb{S}^2$). Point coordinates correspondences between frames are given by the warping function $w : \mathbb{P}^2 \times \mathbb{R}_+ \times \mathbb{SE}(3) \mapsto \mathbb{P}^2$, under observability conditions at different viewpoints. From the spherical normalization $\mathbf{q}_S(\mathbf{p}) = g(\mathbf{p})/||g(\mathbf{p})||$ then $g(\mathbf{p}) = \mathcal{D}(\mathbf{p})\mathbf{q}_S(\mathbf{p})$, with $\mathbf{q}_S \in \mathbb{S}^2$ being the unit vector, the warping is given by:

$$w(\mathbf{p},\mathcal{D}(\mathbf{p}),\mathbf{T}) = \mathbf{q}_S^{-1}\left(\frac{\mathbf{R}(\mathcal{D}(\mathbf{p})\mathbf{q}_S(\mathbf{p})) + \mathbf{t}}{||\mathbf{R}(\mathcal{D}(\mathbf{p})\mathbf{q}_S(\mathbf{p})) + \mathbf{t}||}\right) \quad (1)$$

where $\mathbf{q}_S^{-1}(\bullet)$ is the inverse unit sphere mapping to Cartesian coordinates. The pose $\mathbf{T}(\mathbf{x}) = \exp([\mathbf{x}]_\wedge) \in \mathbb{SE}(3)$ linking two frames (reference and target frames) is defined by the exponential map of the six DOF twist velocities $\mathbf{x} = (\boldsymbol{\upsilon}\delta t, \boldsymbol{\omega}\delta t) \in \mathbb{R}^6$, with

$$[\mathbf{x}]_\wedge = \begin{bmatrix} \mathbf{S}(\boldsymbol{\omega})\delta t & \boldsymbol{\upsilon}\delta t \\ \mathbf{0}_{(1\times 3)} & 0 \end{bmatrix} \in \mathfrak{se}(3) \quad (2)$$

which is the Lie algebra of $\mathbb{SE}(3)$ at the identity element; $\mathbf{S}(\bullet)$ represents the skew symmetric cross product matrix and $\delta t = 1$. Finally, the scene environment model is represented as a subset of georeferenced frames (named as keyframes) in a sparse graph structure (see top fig. 1). Such keyframes are chosen when they provide new information about the scene according to an entropy/MAD criteria as in [KSC13] [GMRD15].

III. Efficient RGB-D Registration

Our adaptive RGB-D registration approach is based on classical direct VO [CMR10] and ICP [GIRL03] strategies. We recall next the main concepts presented in [MFMR16], that are useful for creating the keyframe map representation. The central idea is that the intensity and depth data error terms display different convergence properties for small and large motions. We will explore these complementary aspects, in terms of convergence, by using a modified cost function, where the geometric term prevails in the first coarse iterations, while the intensity data term dominates in the finer increments. One of the main goals of this work is to evaluate the influence of the registration method in the mapping aspect, i.e.,, the graph construction from a sequence.

A. Adaptive Hybrid RGB-D Cost Function

The pose $\hat{\mathbf{T}}\mathbf{T}(\mathbf{x})$ between a frame and the learned model is performed iteratively from a linearised convex cost function of the following photometric and geometric errors

$$e_I(\mathbf{p}, \mathbf{x}) = \mathcal{I}(w(\mathbf{p}, \hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) - \mathcal{I}^*(\mathbf{p}) \quad (3)$$

$$e_D(\mathbf{p}, \mathbf{x}) = \lambda_D(\hat{\mathbf{R}}\mathbf{R}(\mathbf{x})\mathbf{n}^*(\mathbf{p}))^T(g(w(\mathbf{p},\hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) - \hat{\mathbf{T}}\mathbf{T}(\mathbf{x})g^*(\mathbf{p})) \quad (4)$$

Where $\hat{\mathbf{T}}$ is the initial pose guess; $\mathbf{n}^*$ the normal surface vector calculated at the reference frame; $g(\bullet)$ is the inverse projection model (3D point coordinates) as in (1); and λ_D is a tuning parameter for scaling the error terms. The intensity only (RGB) registration method is equivalent to consider $\lambda_D = 0$. Eq. (3) is a classical optical flow constraint equation (OFCE) term (within the hypothesis of Lambertian surfaces) and (4) is equivalent to a flow point-to-plane ICP, both assuming predominant static surfaces. To ensure these assumptions, robust M-estimators (denoted as $\rho(\bullet)$) are coupled to these errors for mitigating outliers influence [Zha95]. This allows to reduce the effects of self occlusions, moving objects, illumination and interpolations errors in the direct estimation. The classic RGB-D registration consists of using jointly (3) and (4) as

$$C(\mathbf{x}) = \min_{\mathbf{x}} \sum_{\mathbf{p}} \rho_I(e_I(\mathbf{p}, \mathbf{x})) + \sum_{\mathbf{p}} \rho_D(e_D(\mathbf{p}, \mathbf{x})) \quad (5)$$

To avoid local minima and for increasing the convergence rate, the optimization of (5) is often done considering multi-resolution Gaussian pyramidal images [TAC11] [KSC13].

An important issue in (5) is the selection of the scaling factor λ_D for ensuring nice convergence properties. Choosing a large λ_D ($\lambda_D >> 1$) in (5) is equivalent to the direct ICP method, while $\lambda_D \approx 0$ corresponds to a classical dense VO. [TAC11] adopts a constant $\lambda_D = median(\mathcal{I}^*)/median(\mathcal{D}^*)$ during all the optimization procedure. We observed that the intensity RGB is flatter than the geometric costs when further from the solution, but locally more precise when near the solution. A better strategy should then to adapt the value of λ_D (i.e. the balance between the RGB and ICP costs) during the optimization. This conclusion leads to the fundamental question of how to identify that the current pose in (5) is in vicinity of the optimal pose and how to define such neighbourhood.

The adopted strategy is to analyse the costs relative behaviour along the optimization steps. The idea is that the relative conditioning number detects when the algorithm is in the vicinity of the solution (i.e. where the ICP cost is less discriminant). We propose an activation function $\mu(\mathbf{x})$ to act in the original cost (5) where the geometric term prevails

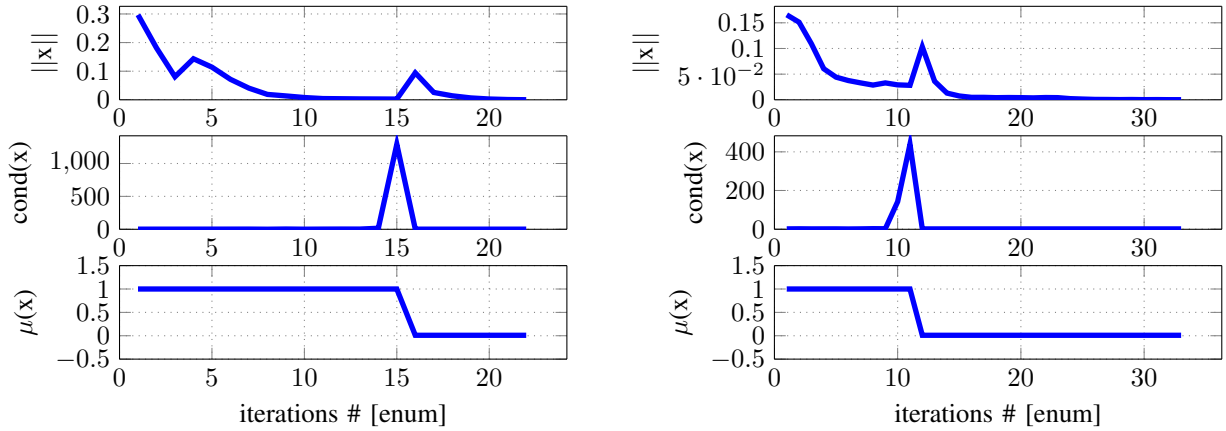

Fig. 2: Activation adaptive function $\mu(\mathbf{x})$ while performing registration in the KITTI outdoor dataset in two different areas (frames' numbers 5 and 100). The left column corresponds to a scene where the convergence is slower (corridor like environment) and the right column is of frames changed predominantly by a rotation. The conditioning criteria (6) is easily detectable for both cases.

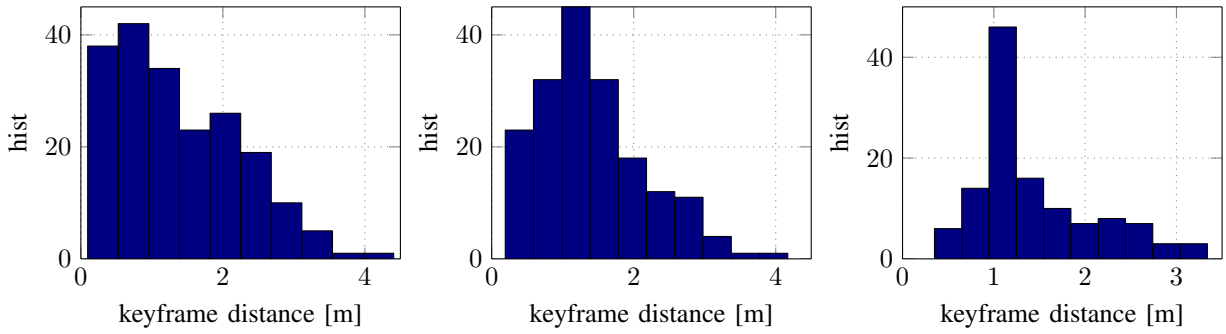

Fig. 3: Distributions of the distances between successive keyframes for the Inria (left), the urban 1 (middle) and the urban 2 (right) sequences.

in the first coarse iterations, while the intensity data term dominates in the finer increments (in the neighbourhood of the solution), as:

$$\mu(\mathbf{x}) = k_1 \mathbf{1}(cond_{\mathbf{x}}(C_I(\mathbf{x}))/cond_{\mathbf{x}}(C_D(\mathbf{x}))) + k_2 \quad (6)$$

where $0 < k_1 \leq k_1 + k_2 < 1$ and $0 \leq k_2 < k_1$, the indicator function $\mathbf{1}(\bullet)$ is zero when $cond_{\mathbf{x}}(C_I(\mathbf{x}))/cond_{\mathbf{x}}(C_D(\mathbf{x})) > \kappa >> 1$ and one otherwise; and

$$cond_{\mathbf{x}}(C(\mathbf{x})) = \left| \frac{C(\mathbf{x}_0 \circ \mathbf{x}) - C(\mathbf{x}_0)}{C(\mathbf{x}_0)} \right| / \frac{||\mathbf{x}||}{||\mathbf{x}_0||} \quad (7)$$

being an estimate of the relative condition number of the RGB (C_I) and ICP (C_D) cost functions, with $\mathbf{x}_0 = [(\log(\hat{\mathbf{T}}))]^{-1}_{\wedge}$ as in (2) and $\circ$ is the additive Lie algebra action. For simplicity, we assume $k_2 = 0$ due to the opposite convergence properties of the cost terms near the solution. We show in fig. 2 typical convergence curves using frames from the KITTI VO/SLAM sequence 00 [GLU12].

The respective modified cost function is designed with $\lambda_D = 1$ in a joint adaptive RGB-ICP cost

$$\tilde{C}(\mathbf{x}) = (1 - \mu(\mathbf{x})) \sum_{\mathbf{p}} \rho_I(e_I(\mathbf{p}, \mathbf{x})) + \\ + \mu(\mathbf{x}) \sum_{\mathbf{p}} \rho_D(e_D(\mathbf{p}, \mathbf{x})) \quad (8)$$

To further increase the convergence, the cost (8) is embedded into a multi-resolution framework. We begin with the smallest resolution (pyramid at level n) to the bigger resolution (pyramid level 1). Finally, the Huber robust function is applied in ρ_I, ρ_D for the first iterations and we switch to Tukey when near the solution [Zha95]. The respective Jacobians and details about the optimization are given in [MFMR16].

IV. Localization and Mapping Experiments

The experiments are performed in outdoor real scenes using a spherical RGB-D acquisition sensor. The sensor was mounted on a typical non-holonomic vehicle. The depth was computed by spherical stereo using ELAS [GRU10]. The multi-resolution framework used a Gaussian pyramid of four levels (the higher the level, the smaller the image resolution is). The maximum number of iterations was of 50 at each

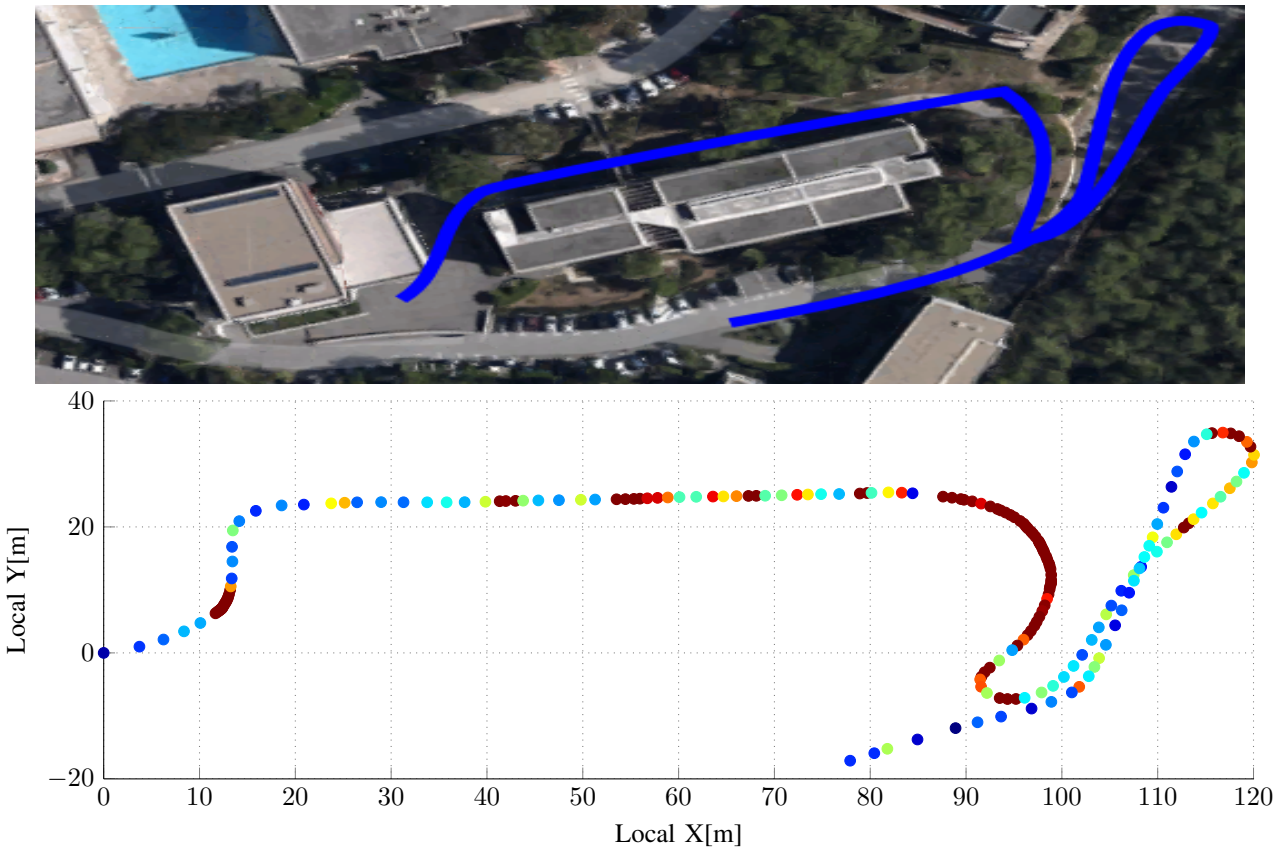

Fig. 4: Outdoor mapping in a semi-urban area. The bird's-eye view of the approximative real trajectory (in blue) and the estimated trajectory using the adaptive formulation (bottom figure). The color along the trajectory indicates the density of keyframes.

pyramid level. The iterative pose estimation algorithm is said to have converged (either to a global or to a local minimum) when the norm of pose increments $\mathbf{x}$ is bellow a fixed threshold in successive iterations (10^{-5} for the rotation and 10^{-3} for translation). A new keyframe is selected when one the following conditions is reached: i) the registration takes the maximum number of iterations in the highest resolution; ii) the rotation and translation relative norms are bigger than 45 degrees and 1 meter respectively and; ii) the MAD of the intensity error between the frames is bigger than 15. Lastly, the parameters of the activation function (6) were $k_1 = 0.95$ and the relative conditioning $\kappa = 5$.

At first, we evaluate the localization performance of the registration technique in controlled conditions using the sequence 00 of the KITTI SLAM/VO benchmark [GRU10]. The convergence was achieved even in cases considering translations and rotations of around 2 meters and 15 degrees. Besides the advantages of higher accuracy and robustness, we remark that there is also an advantage on the computational cost during the localization task. Convergence is reached with a reduced number of iterations, and thus, the time required to register a pair of frames is at least as twice as fast in average with respect to a classic RGB-D formulation [TAC11].

TABLE I: Spatial keyframe density distribution statistics (unities in meters): mean, median, minimum/maximum and standard deviation (Std)/median absolute deviation (MAD).

	Mean	*Median*	*Min/Max*	*Std/MAD*
Inria sequence	1.4	1.2	0.1/4.4	0.89/0.94
Urban sequence 1	1.5	1.2	0.2/4.2	0.75/0.68
Urban sequence 2	1.2	1.2	0.1/3.3	0.60/0.59

We then realize the localization with the keyframe selection (mapping) using the spherical stereo sensor. These experiments were performed in three different outdoor urban areas (see figs. 4, 5 and 6). The adaptive formulation allowed a consistent pose estimates in all cases (see the trajectories in fig. 4). An important property of the mapping in large scales is the compactness, i.e, how sparse is the topo-metric graph. We present in Table I and in fig. 3 some metrics of the admissible distances between the selected keyframes.

The density of retained keyframes along the trajectory is encoded by the color (brighter regions indicate higher keyframe density, i.e. smaller distance) in figs. 4, 5, 6. We can clearly identify the regions with higher density as the areas of turns with partial occlusion of the scene or in cases of failure of the Lambertian property. In the blue sections, encoding regions with invariant viewing conditions (e.g. with a predominantly convex geometry and verifying

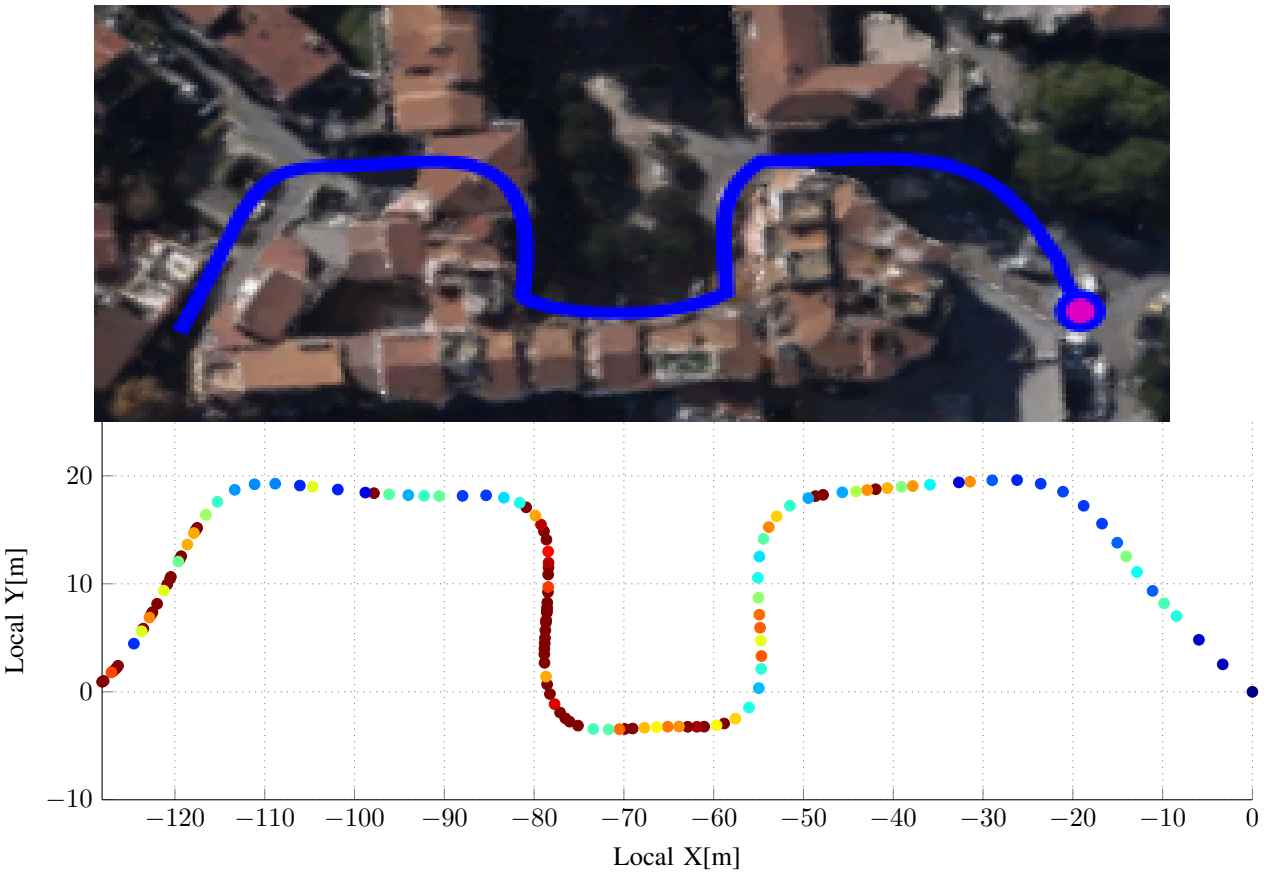

Fig. 5: Outdoor mapping example in a urban area. The bird's-eye view of the approximative real trajectory is shown in the top image (in blue) and the estimated trajectory using the adaptive formulation in the bottom plot. The color along the trajectory indicates the density of keyframes.

the Lambertian hypothesis), the sparsity of the keyframes was of up to 4.4 meters (see Table I), which turns to be more than three times bigger than with VO only based registration techniques. We reinforce that the sparsity of the map might be adapted to the capacities of the posterior registration algorithm to explore it. Using the adaptive RGB-D formulation, only 3.5% (207/5200) of the frames were used in the model (keyframes) in the Inria sequence and only 8% (183/2100) and 6.8% (102/1500) in the urban ones.

V. Conclusions

In this paper, we presented an efficient RGB-D registration approach in the context of large inter-frame displacements and its application to outdoor scene mapping. This is of fundamental interest in large scale navigation/mapping scenarios and in compact mapping since it allows to create a sparser local representation whilst maintaining a topological structure at large-scale that is accurate enough to ensure the convergence of a task in the neighbourhood of the scene model. The technique exploits adaptively the photometric and geometric error terms based on their convergence characteristics in a multi-resolution framework. Despite its simplicity, this approach was capable of dealing with large rotations, occlusions and moving objects presented in real outdoor scenarios.

Future directions include: (i) the optimal positioning of keyframes (e.g. by efficient space partitioning techniques as Voronoi tessellations); (ii) the prediction of scene occlusions by analysing the scene geometry convexity (e.g. by observing the properties of local piece-wise patches); and (iii) the adaptation of the keyframe selection/density from experience, e.g., considering convergence bounds from teach and repeat paradigm [CTG+15].

References

[CMR10] A. Comport, E. Malis, and P. Rives. Real-time quadrifocal visual odometry. *IJRR*, 29, 2010.

[CTG+15] W. Churchill, C. Tong, C. Gurau, I. Posner, and P. Newman. Know your limits: Embedding localiser performance models in teach and repeat maps. In *IEEE ICRA*, 2015.

[FB10] P. Furgale and T. Barfoot. Visual teach and repeat for long-range rover autonomy. *JFR*, 27, 2010.

[GIRL03] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy. Geometrically stable sampling for the icp algorithm. In *3DIM*, 2003.

[GLU12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE CVPR*, 2012.

[GMRD15] T. Gokhool, R. Martins, P. Rives, and N. Despre. A compact spherical RGBD keyframe-based representation. In *IEEE ICRA*, 2015.

[GRU10] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010.

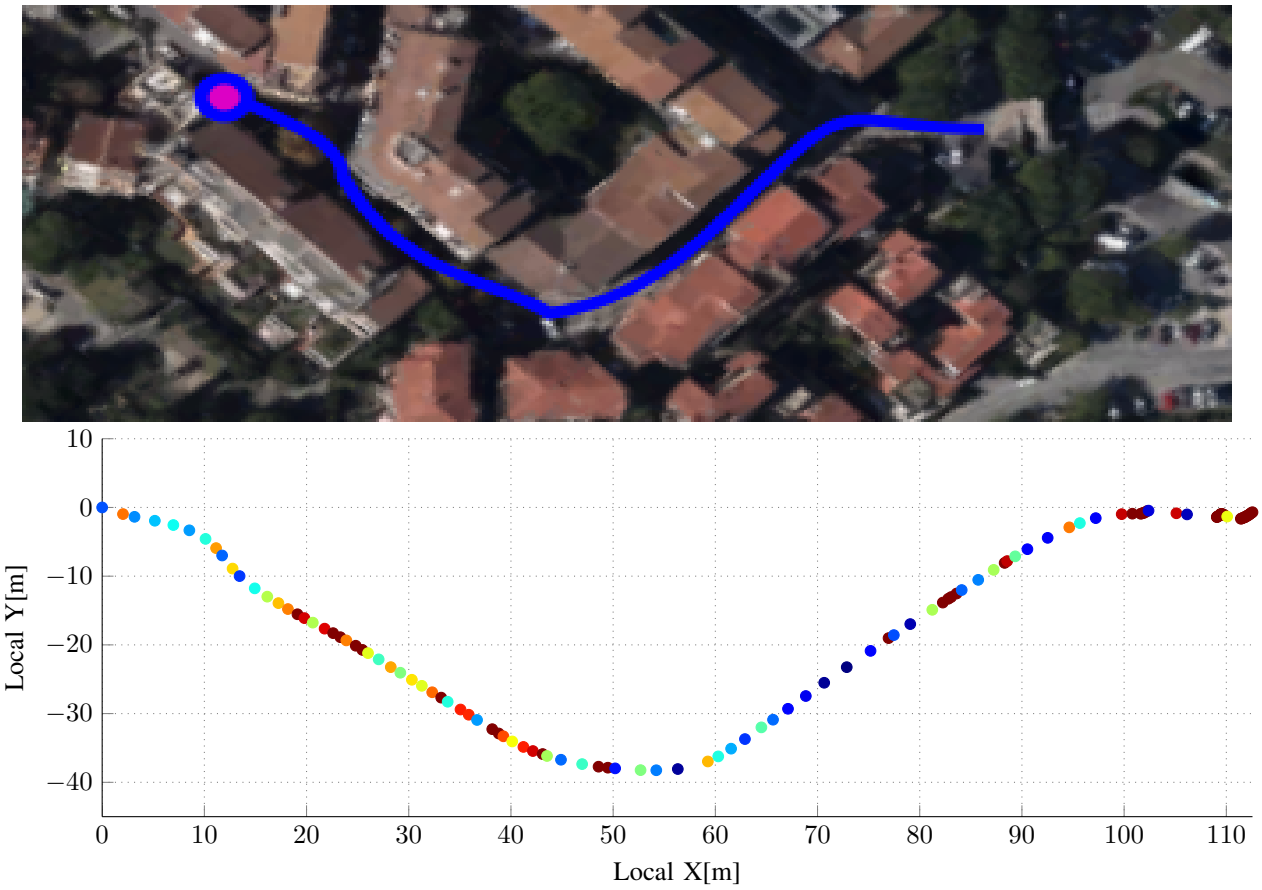

Fig. 6: Outdoor mapping example in a urban area. The bird's-eye view of the approximative real trajectory is shown in the top image (in blue) and the estimated trajectory using the adaptive formulation in the bottom plot. The color along the trajectory indicates the density of keyframes.

[How08] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, Nice, France, 2008.

[KA08] K. Konolige and M. Agrawal. Frameslam: from bundle adjustment to realtime visual mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, 2008.

[KSC13] C. Kerl, J. Sturm, and D. Cremers. Dense visual SLAM for RGB-D cameras. In *IEEE IROS*, 2013.

[MC13] M. Meilland and A. Comport. On unifying key-frame and voxel-based dense visual slam at large scales. In *IEEE IROS*, 2013.

[MCR15] M. Meilland, A. Comport, and P. Rives. Dense omnidirectional RGB-D mapping of large-scale outdoor environments for real-time localization and autonomous navigation. *JFR*, 32(4), 2015.

[MFMR15] R. Martins, E. Fernandez-Moral, and P. Rives. Dense accurate urban mapping from spherical RGB-D images. In *IEEE IROS*, 2015.

[MFMR16] R. Martins, E. Fernandez-Moral, and P. Rives. Adaptive direct RGB-D registration and mapping for large motions. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[NNB04] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[TAC11] T.Tykkala, C. Audras, and A. Comport. Direct iterative closest point for real-time visual odometry. In *ICCV Workshops*, 2011.

[Zha95] Z. Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. Technical Report 2676, Inria, 1995.